Published online 10 November 2011 Nucleic Acids Research, 2012, Vol. 40, Database issue D935–D939
doi:10.1093/nar/gkr996

SNPeffect 4.0: on-line prediction of molecular and structural effects of protein-coding variants

Greet De Baets[1,2,3], Joost Van Durme[1,2,3], Joke Reumers[4,5], Sebastian Maurer-Stroh[6], Peter Vanhee[1,3], Joaquin Dopazo[7,8], Joost Schymkowitz[1,2,*] and Frederic Rousseau[1,2,*]

[1]VIB Switch Laboratory, 3000 Leuven, [2]Department of Molecular Cell Biology, University of Leuven, [3]Vrije Universiteit Brussel, [4]Vesalius Research Center, VIB, 3000 Leuven, [5]Vesalius Research Center, University of Leuven, Belgium, [6]Bioinformatics Institute (BII), Agency for Science, Technology and Research (A*STAR), Singapore, [7]Department of Bioinformatics, Centro de Investigación Príncipe Felipe (CIPF) and [8]Functional Genomics Node (INB) at CIPF, Valencia 46013, Spain

Received September 14, 2011; Revised October 17, 2011; Accepted October 18, 2011

ABSTRACT

Single nucleotide variants (SNVs) are, together with copy number variation, the primary source of variation in the human genome and are associated with phenotypic variation such as altered response to drug treatment and susceptibility to disease. Linking structural effects of non-synonymous SNVs to functional outcomes is a major issue in structural bioinformatics. The SNPeffect database (http://snpeffect.switchlab.org) uses sequence- and structure-based bioinformatics tools to predict the effect of protein-coding SNVs on the structural phenotype of proteins. It integrates aggregation prediction (TANGO), amyloid prediction (WALTZ), chaperone-binding prediction (LIMBO) and protein stability analysis (FoldX) for structural phenotyping. Additionally, SNPeffect holds information on affected catalytic sites and a number of post-translational modifications. The database contains all known human protein variants from UniProt, but users can now also submit custom protein variants for a SNPeffect analysis, including automated structure modeling. The new meta-analysis application allows plotting correlations between phenotypic features for a user-selected set of variants.

INTRODUCTION

Human next-generation sequencing projects currently generate millions of previously unknown single nucleotide variants (SNVs) (1). On average, every newly sequenced genome generates about 300 000 novel SNVs (2). Although it is quite straightforward to annotate these SNVs according to their genomic location (coding, non-coding and regulatory regions), and for coding SNVs to denote their effect on the translated protein (synonymous or non-synonymous), predicting the detailed effect of a coding mutation on the structure and function of a protein is a largely unsolved problem. As these variants can influence drug selection, dosing and adverse effects (3), it is recognized that this genetic information is of great importance for drug development in general (4) and crucial for personalized medicine (5). Most current approaches classify SNVs into neutral or deleterious variants by using either conservation based measures (6) or by using a combination of conservation scores and structural features (7–9). Tools for predicting stability changes upon mutation have also been developed (10,11), however these do not use explicit stability predictions based on a high-resolution structure but rather depend on black-box predictions using intelligent machine-learning approaches such as support-vector machines or neural networks.

Coding non-synonymous SNVs can affect protein structure and function to various degrees (12,13). Although predicting neutral or fully disruptive variants is relatively easy, a large portion of variants will result in more subtle intermediate phenotypic effects that are much more challenging to predict.

To tackle this challenge web servers such as PolyPhen (9) and HOPE (8), for example, base their predictions on a statistical analysis of protein structures extrapolated to the protein under study and do currently not provide quantitative free energy changes of point mutations. SNPeffect on the other hand uses the FoldX (14) force field and aims

*To whom correspondence should be addressed. Tel: +32 16 37 25 70; Fax: +32 16 37 25 71; Email: Joost.schymkowitz@switch.vib-kuleuven.be
Correspondence may also be addressed to Frederic Rousseau. Tel: +32 16 37 25 70; Fax: +32 16 37 25 71; Email: Frederic.rousseau@switch.vib-kuleuven.be

The authors wish it to be known that, in their opinion, the first two authors should be regarded as joint First Authors.

© The Author(s) 2011. Published by Oxford University Press.
This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (http://creativecommons.org/licenses/by-nc/3.0), which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015

at calculating realistic free-energy changes upon mutation ($\Delta\Delta G$), thereby providing high-accuracy protein stability information. As structure quality is crucial for the accuracy of $\Delta\Delta G$ predictions using FoldX we currently do not model structures with <90% sequence identity to the modeling template structure. As a result the structural coverage of SNPeffect is somewhat lower than that of PolyPhen or HOPE. However, by integrating several in-house developed structural bioinformatics tools designed to quantify protein misfolding (FoldX), protein aggregation [TANGO (15) and WALTZ (16)] and chaperone interaction [LIMBO (17)], SNPeffect was developed with the specific aim of mapping the effect of SNVs on the protein homeostasis landscape. i.e. the ability of a cell to maintain appropriate concentrations of properly folded proteins in the correct cellular compartment (18). Currently SNPeffect provides pre-calculated mutant analyses for more than 60 000 human coding protein variants, benefiting the speed of information retrieval, but it also allows calculation of custom mutant sets. Finally SNPeffect provides features for meta-analysis of selected data sets allowing to analyze the proteostatic landscape of a given protein or protein family for example.

SNPeffect PIPELINE FOR MOLECULAR PHENOTYPING OF HUMAN PROTEIN VARIANTS

The raw data source of the SNPeffect database consists of the UniProt human variation database (http://www.uniprot.org/docs/humsavar), containing single amino acid polymorphisms, classified either as disease mutations, polymorphisms or yet unclassified mutations. SNPeffect predicts the impact of these variants on (i) protein aggregation and amyloid formation (TANGO and WALTZ, respectively), (ii) chaperone binding (LIMBO) and (iii) structural stability (FoldX). The availability of a crystal structure with a minimal resolution of 4 Å is required to accurately analyze the effect on protein stability with FoldX. If an exact structural match is not found, homologous structures with no <90% sequence identity are considered as template structures to build a homology model of the original sequence with FoldX. The stability analysis is then applied to this model.

Furthermore, SNPeffect holds annotations on functional sites, structural features, domain information, cellular processing and post-translational modifications for each variant.

The effect on functional sites and structural features is analyzed by investigating several properties of the position of the mutation. Data from the Catalytic Site Atlas is parsed to analyze whether the residue is part of the active site (19). Secondary structure information is generated by FoldX and transmembrane topology (extracellular, intracellular, transmembrane) is predicted by TMHMM (20). Domain information is provided by SMART (21) and PFAM (22). PSORT (23) provides a prediction on the sub-cellular localization. SNPeffect also maps changes in post-translational lipid anchor attachment and the peroxisomal targeting signal PTS1 (24). Lipid attachment predictions include myristoylation (25,26), farnesylation (26), GPI-anchor attachment (27) and type-1 and type-2 geranylgeranylation (26).

All entries are additionally linked to the OMIM genetic disorder database (Online Mendelian Inheritance in Man, OMIM. McKusick-Nathans Institute of Genetic Medicine, Johns Hopkins University (Baltimore, MD), 2011. World Wide Web URL: http://omim.org/) and the Gene Ontology database (28).

SNPeffect DATABASE

SNPeffect currently contains data on 63 410 human non-synonymous SNVs. Automatic updates from the UniProt human variation database are scheduled every 6 months.

The database interface (Figure 1, left) allows users to search SNVs by filtering on molecular phenotypic effects, mutation type, disease, UniProt identifier, dbSNP identifier and gene name. Molecular phenotypic effects include changes in aggregation tendency (dTANGO), amyloid formation (dWALTZ), chaperone binding (dLIMBO) and structural stability change upon mutation (ddG). Applying the filter settings results in a set of variants that can be analyzed in a protein-centered or variant-centered view (Figure 1, right).

This SNPeffect update focuses primarily on the scientist user's ability to quickly retrieve and rapidly analyze the effect of protein variants. Moreover, the wild-type protein of each variant is also fully analyzed and directly linked from the variant webpage. The effects are visualized by self-explanatory barplots and histograms. Structural data is retrieved from the Protein Data Bank (PDB) (29). When an exact match to the wild-type sequence is not found, a homology model is built from a template structure that has at least 90% sequence identity to the original sequence. If structural information is retrieved, we offer visualization of both the wild-type and mutant residue environment in the protein structure. Additionally, every phenotypic analysis is accompanied by a graphical and textual comparison to the wild-type protein. Figure 1 illustrates the summary of a variant that meets the criteria set in the filter.

META-ANALYSIS

A new feature in SNPeffect 4.0 is the ability to analyze and plot phenotypic features of a specific subset (or all) of the SNPeffect database. The meta-analysis tool enables scientists to carry out large-scale data mining of the specified data and visualize the results in a graphical plot. The data set of variants is primarily chosen on disease associations and the mutation type. Mutation types include disease, polymorphism and unclassified. An additional filter can be applied to limit the results by one or more disease terms that are selected from a list or specified by keywords. SNPeffect will then search for all variants of the selected type and retrieve those that are linked to the selected disease(s). For the disease type, these are solely the mutations annotated with that disease. For the polymorphisms and unclassifieds, SNPeffect retrieves all of

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015

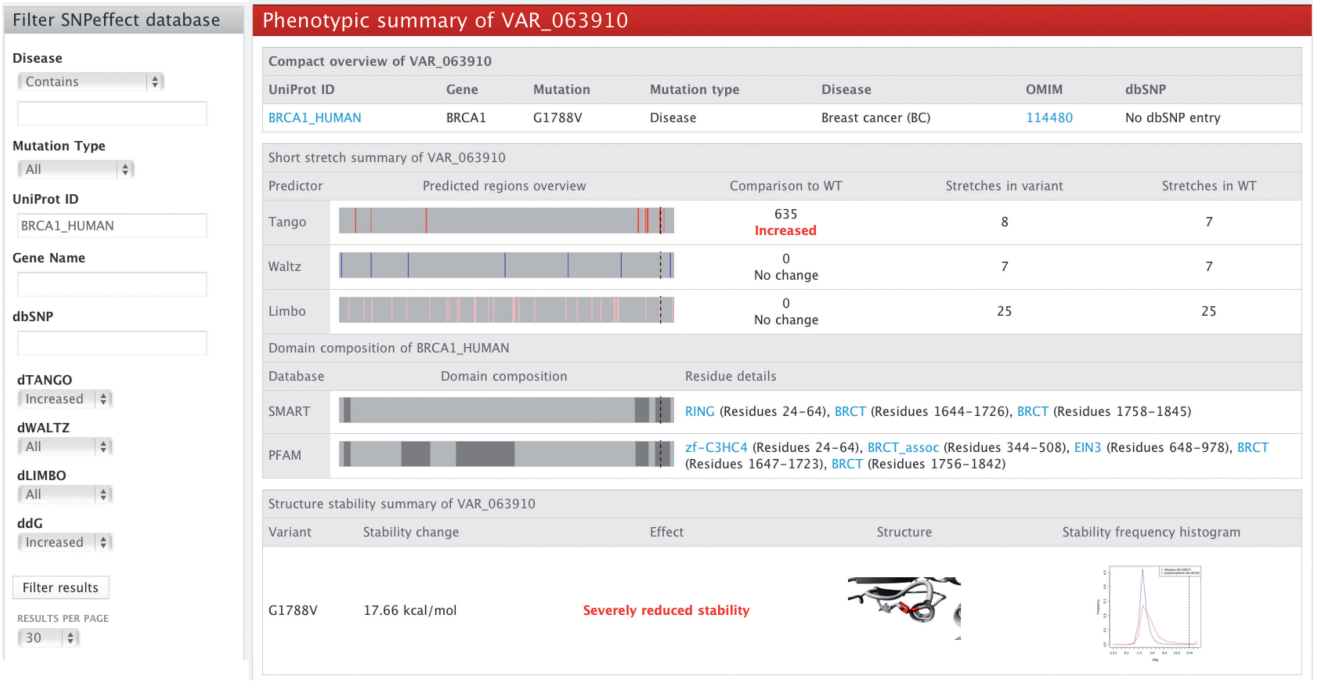

Figure 1. Phenotypic summary of a variant. In the form on the left, the filter settings are selected. The webpage on the right shows a variant that meets the filter criteria and displays summarized information on the effect on aggregation tendency, amyloid propensity, chaperone binding and structural stability, as well as domain annotation from the SMART and PFAM databases. Below the phenotypic summary section, detailed information from all predictors can be consulted for an even deeper variant analysis.

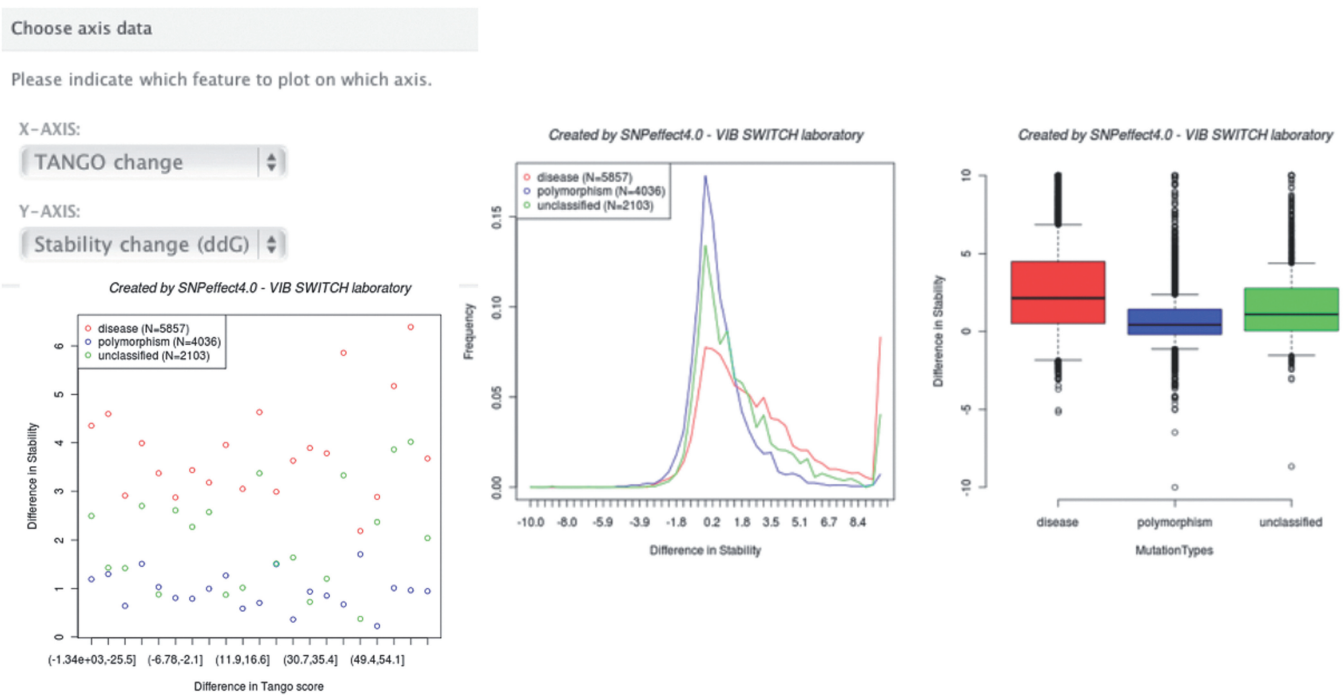

Figure 2. Overview of the meta-analysis tool. The left top shows the form to specify which phenotypic features to plot. The bottom images show (from left to right) a scatter plot, a histogram and a boxplot for the selected data set.

these variants from proteins associated with the selected disease(s). Next, the two phenotypic effects that will be analyzed and plotted can be specified (dTANGO, dWALTZ, dLIMBO or ddG) (Figure 2). For example, one can create an aggregation/stability feature plot of a set of variants to correlate aggregation changes with

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015

stability changes. If the number of hits for one of the mutation types exceeds 500, the average Y is plotted for each X bin, to keep the plots clear and readable. The meta-analysis tool converts phenotypic features of a selected set of variants to comprehensible scatter plots, boxplots and frequency plots (Figure 2).

JOB SUBMISSION

Novel to previous versions of SNPeffect (30–32), the current version includes a data submission framework that allows submitting (human or non-human) custom single protein variants for a detailed SNPeffect analysis including TANGO, WALTZ, LIMBO and FoldX. Possible input types are UniProt ID, FASTA sequence, PDB ID, or an uploaded PDB file. If only sequence information but no structural information is provided, SNPeffect will search the PDB for a matching structure to complete the stability analysis with FoldX. When an exact match is not found, a *homology* filter allows setting the minimum percent sequence identity a structural homolog template should have to build a homology model. The effect on structural stability is then determined by analyzing the homology model. Users receive an e-mail notification when the analysis has finished and can download the results from their SNPeffect account. The results include a PDF file with the complete phenotypic SNPeffect analysis. This file contains figures and extensive life scientist-friendly text reports with comparison to the wild-type protein. All separate figure files are also available and free to use.

SUMMARY

SNPeffect 4.0 offers a detailed and comprehensible molecular and structural phenotypic analysis of all known human protein variants. Major phenotypic features such as aggregation propensity prediction, stability analysis, structural features, post-translational modification and cellular localization are intelligibly visualized and explained for each variant. The meta-analysis tool allows plotting correlations between phenotypic effects concerning a specified set of variants. Custom protein variants can now be submitted for a detailed SNPeffect analysis, including automated structure modeling. SNPeffect 4.0 is available at http://snpeffect.switchlab.org

FUNDING

Interuniversity Attraction Poles (IAP Network 6/43) of the Belgian Federal Science Policy Office (BelSPo) (VIB Switch laboratory); Flanders Institute for Science and Technology (IWT) (to G.D.B.); Fund for Scientific Research (FWO), Flanders (to J.V.D.); MICINN projects BIO2008-04212, and RD06/0020/1019 (RTICC, ISCIII) (Dopazo lab, partial) GVA-FEDER (PROMETEO/2010/001, partial). The CIBER de Enfermedades Raras is an initiative of the ISCIII, MICINN. Funding for open access charge: VIB.

Conflict of interest statement. None declared.

REFERENCES

1. Alkan,C., Coe,B.P. and Eichler,E.E. (2011) Genome structural variation discovery and genotyping. *Nat. Rev. Genet.*, **12**, 363–376.
2. Collins,F.S., Guyer,M.S. and Charkravarti,A. (1997) Variations on a theme: cataloging human DNA sequence variation. *Science*, **278**, 1580–1581.
3. Giacomini,K.M., Brett,C.M., Altman,R.B., Benowitz,N.L., Dolan,M.E., Flockhart,D.A., Johnson,J.A., Hayes,D.F., Klein,T., Krauss,R.M. *et al.* (2007) The pharmacogenetics research network: from SNP discovery to clinical drug response. *Clin. Pharmacol. Ther.*, **81**, 328–345.
4. Foot,E., Kleyn,D. and Palmer Foster,E. (2010) Pharmacogenetics–pivotal to the future of the biopharmaceutical industry. *Drug Discov. Today*, **15**, 325–327.
5. Fernald,G.H., Capriotti,E., Daneshjou,R., Karczewski,K.J. and Altman,R.B. (2011) Bioinformatics challenges for personalized medicine. *Bioinformatics*, **27**, 1741–1748.
6. Kumar,P., Henikoff,S. and Ng,P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.
7. Sunyaev,S., Ramensky,V. and Bork,P. (2000) Towards a structural basis of human non-synonymous single nucleotide polymorphisms. *Trends Genet.*, **16**, 198–200.
8. Venselaar,H., Te Beek,T.A., Kuipers,R.K., Hekkelman,M.L. and Vriend,G. (2010) Protein structure analysis of mutations causing inheritable diseases. An e-Science approach with life scientist friendly interfaces. *BMC Bioinformatics*, **11**, 548.
9. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
10. Bromberg,Y. and Rost,B. (2007) SNAP: predict effect of non-synonymous polymorphisms on function. *Nucleic Acids Res.*, **35**, 3823–3835.
11. Capriotti,E., Fariselli,P. and Casadio,R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, **33**, W306–W310.
12. Hartl,F.U., Bracher,A. and Hayer-Hartl,M. (2011) Molecular chaperones in protein folding and proteostasis. *Nature*, **475**, 324–332.
13. Tokuriki,N. and Tawfik,D.S. (2009) Chaperonin overexpression promotes genetic variation and enzyme evolution. *Nature*, **459**, 668–673.
14. Schymkowitz,J., Borg,J., Stricher,F., Nys,R., Rousseau,F. and Serrano,L. (2005) The FoldX web server: an online force field. *Nucleic Acids Res.*, **33**, W382–W388.
15. Fernandez-Escamilla,A.M., Rousseau,F., Schymkowitz,J. and Serrano,L. (2004) Prediction of sequence-dependent and mutational effects on the aggregation of peptides and proteins. *Nat. Biotechnol.*, **22**, 1302–1306.
16. Maurer-Stroh,S., Debulpaep,M., Kuemmerer,N., Lopez de la Paz,M., Martins,I.C., Reumers,J., Morris,K.L., Copland,A., Serpell,L., Serrano,L. *et al.* (2010) Exploring the sequence determinants of amyloid structure using position-specific scoring matrices. *Nat. Methods*, **7**, 237–242.
17. Van Durme,J., Maurer-Stroh,S., Gallardo,R., Wilkinson,H., Rousseau,F. and Schymkowitz,J. (2009) Accurate prediction of DnaK-peptide binding via homology modelling and experimental data. *PLoS Comput. Biol.*, **5**, e1000475.
18. Powers,E.T., Morimoto,R.I., Dillin,A., Kelly,J.W. and Balch,W.E. (2009) Biological and chemical approaches to diseases of proteostasis deficiency. *Annu. Rev. Biochem.*, **78**, 959–991.
19. Torrance,J.W., Bartlett,G.J., Porter,C.T. and Thornton,J.M. (2005) Using a library of structural templates to recognise catalytic sites and explore their evolution in homologous families. *J. Mol. Biol.*, **347**, 565–581.

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015

20. Krogh,A., Larsson,B., von Heijne,G. and Sonnhammer,E.L. (2001) Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. *J. Mol. Biol.*, **305**, 567–580.
21. Letunic,I., Doerks,T. and Bork,P. (2009) SMART 6: recent updates and new developments. *Nucleic Acids Res.*, **37**, D229–D232.
22. Finn,R.D., Mistry,J., Tate,J., Coggill,P., Heger,A., Pollington,J.E., Gavin,O.L., Gunasekaran,P., Ceric,G., Forslund,K. *et al.* (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.
23. Nakai,K. and Horton,P. (1999) PSORT: a program for detecting sorting signals in proteins and predicting their subcellular localization. *Trends Biochem. Sci.*, **24**, 34–36.
24. Neuberger,G., Maurer-Stroh,S., Eisenhaber,B., Hartig,A. and Eisenhaber,F. (2003) Prediction of peroxisomal targeting signal 1 containing proteins from amino acid sequence. *J. Mol. Biol.*, **328**, 581–592.
25. Maurer-Stroh,S., Eisenhaber,B. and Eisenhaber,F. (2002) N-terminal N-myristoylation of proteins: prediction of substrate proteins from amino acid sequence. *J. Mol. Biol.*, **317**, 541–557.
26. Maurer-Stroh,S. and Eisenhaber,F. (2005) Refinement and prediction of protein prenylation motifs. *Genome Biol.*, **6**, R55.
27. Eisenhaber,B., Bork,P. and Eisenhaber,F. (1999) Prediction of potential GPI-modification sites in proprotein sequences. *J. Mol. Biol.*, **292**, 741–758.
28. Ashburner,M., Ball,C.A., Blake,J.A., Botstein,D., Butler,H., Cherry,J.M., Davis,A.P., Dolinski,K., Dwight,S.S., Eppig,J.T. *et al.* (2000) Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet*, **25**, 25–29.
29. Berman,H.M., Westbrook,J., Feng,Z., Gilliland,G., Bhat,T.N., Weissig,H., Shindyalov,I.N. and Bourne,P.E. (2000) The Protein Data Bank. *Nucleic Acids Res.*, **28**, 235–242.
30. Reumers,J., Maurer-Stroh,S., Schymkowitz,J. and Rousseau,F. (2006) SNPeffect v2.0: a new step in investigating the molecular phenotypic effects of human non-synonymous SNPs. *Bioinformatics*, **22**, 2183–2185.
31. Reumers,J., Schymkowitz,J., Ferkinghoff-Borg,J., Stricher,F., Serrano,L. and Rousseau,F. (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Res.*, **33**, D527–D532.
32. Reumers,J., Conde,L., Medina,I., Maurer-Stroh,S., Van Durme,J., Dopazo,J., Rousseau,F. and Schymkowitz,J. (2008) Joint annotation of coding and non-coding single nucleotide polymorphisms and mutations in the SNPeffect and PupaSuite databases. *Nucleic Acids Res.*, **36**, D825–D829.

Downloaded from http://nar.oxfordjournals.org/ at OHIO STATE UNIVERSITY LIBRARIES on December 9, 2015